# 多层次研究的数据聚合适当性检验：文献评价与关键问题试解

朱海腾

（陆军炮兵防空兵学院军政基础系，合肥 230031）

**摘　要** 组织管理领域的多层次研究经常需要测量共享单位特性构念，常用方法是将单位内若干个体成员的评分聚合到单位层次，确保聚合后的分数具有充分代表性的统计前提是通过聚合适当性检验。聚合适当性检验的常用指标是组内一致性 $r_{WG}$ 和组内信度 ICC(1)、ICC(2)，但目前学界对于这两类指标何者更优、$r_{WG}$ 的原分布选择和数据清理、各指标的划界值等关键问题存在诸多争议。为此，首先对国内 9 份管理学、心理学期刊 2014 年以来发表的 166 篇包含聚合适当性检验的论文进行内容分析，并以 *Journal of Applied Psychology* 上的 85 篇论文为对比，查明常规实践中的共性问题，进而提出实践建议：（1）明确功能定位，将 $r_{WG}$ 作为聚合适当性指标，ICC(1)和 ICC(2)分别作为效度、信度指标。（2）计算 $r_{WG}$ 时审慎选择原分布，排除组内一致性过低的组。（3）为各指标设定更加合理、有适度灵活性的划界值，停止使用武断、粗糙的经验标准。最后，强调研究者在模型构建和聚合决策中应加强理论考量，避免片面依赖统计检验结果。

**关键词** 多层次研究；共享单位特性；聚合；组内一致性；组内信度

**通讯作者：**朱海腾，E-mail: prettypig1990@sina.com



## 1 引言

多层次组织管理研究经常需要测量处于团体或组织水平的高层次构念。实证研究中最常见的高层次构念是共享单位特性（shared unit property）构念[1]，如团队凝聚力、团队效能，测量这类构念时，通常根据合成模型（composition model）的思想，由团体内的若干成员分别做出评定，取成员评分的均值作为高层次构念的代理值（proxy），这就是数据聚合（aggregation）(Chan, 1998)。为保证聚合后的构念能够代表全体成员的“共享”知觉，需要满足一个统计前提，即团队成员的评分有足够的相似性(Cohen, Doveh, & Eick, 2001; Klein & Kozlowski, 2000; 林钲棽, 彭台光, 2006)，评估数据能否达到这一“门槛”的方法被称为数据聚合适当性检验。

数据聚合适当性检验有两条独特而又互补的路径(Kozlowski & Klein, 2000; LeBreton & Senter, 2008; Shen, 2016; 张志学, 2010)。一是组内一致性/共识（within-group agreement/consensus）检验，衡量多个评定者对某一构念的评分的绝对一致性，即评分是否完全相等，常用指标是 $r_{WG}$[2]。充分的组内一致性既是共享特性构念的构成要素，也是构念效度的证据(Bliese, 2000; James, 1982; Klein, Conn,

[1] 与之相近的一个概念是“情境变量”或“脉络变量”（contextual variable），指的也是由个体层次的观测数据聚合而来的高层次构念，但情境变量不仅包括共享单位特性构念，还包括单纯反映群体特征、不要求组内同质性的生成性（formative）构念，如将学校中所有学生的社会经济地位取均值形成学校平均社会经济地位，它不需要以所有学生有相近的社会经济地位为前提(参见 方杰, 邱皓政, 张敏强, 2011; 于海波, 方俐洛, 凌文辁, 2004)。这类构念不在本文讨论范围之内，后文的“高层次构念”亦特指共享单位特性构念。

[2] 评定项目只有 1 个时记作 $r_{WG(1)}$，有 $J$ 个平行项目时记作 $r_{WG(J)}$。为行文方便，后文统一使用 $r_{WG}$。

Smith, & Sorra, 2001)。二是组内信度（within-group reliability）检验，衡量多个评定者评分的相对一致性，即评分的相对等级是否相同，而不是绝对分数是否相等，常用指标包括组内相关系数 ICC(1)、ICC(2)、方差分析的 $\eta^2$ 等。这两“族”指标分别触及聚合适当性的不同侧面，往往在研究中结合使用。随着多层次研究成为组织管理研究的主流范式，高层次构念的测量和数据聚合一度成为热门方法学议题，但从当前的研究进展来看，数据聚合适当性检验还面临一些悬而未决的难题，其中最关键的问题有三个，分别与指标选择、指标计算和结果解释有关。

第一，关于组内一致性和组内信度何者为数据聚合的“黄金标准”。组内一致性和组内信度分别关注组内变异和组间变异，其理论基础、侧重点、计算和解释方法各有不同，究竟哪一指标可以为聚合决策提供更有价值的信息，一直吸引着研究者的兴趣。George 和 James (1993)曾明确指出，聚合只有两个必要条件，一是在理论上证明某构念应定位于团体层次，二是在统计上证明组内成员的评分有足够的共识。换言之，聚合适当性与组间差异无关，组内一致性才是首要甚至唯一标准(Newman & Sin, 2020)。不过，组内一致性的指标 $r_{WG}$ 有很多局限性，如易受题项数量和组内人数（group size）影响、原分布（null distribution）设定不当导致估计偏差等(Brown & Hauenstein, 2005; O’Neill, 2017)，因此多数研究会同时报告组内信度以期弥补这些缺点。还有研究(Woehr, Loignon, Schmidt, Loughry, & Ohland, 2015)构造了模拟数据，发现 $r_{WG}$ 辨别“伪一致性”的能力不及 ICC(1)和 ICC(2)，建议以组内信度作为聚合的主要标准。虽然组内一致性和组内信度在实际应用中可以并行不悖，但研究者大多只是简单罗列结果，未能对二者功能和角色上的差异进行细致考察。

第二，关于 $r_{WG}$ 的计算和使用，主要涉及原分布选择和数据清理问题。$r_{WG}$ 是一个标准化度量(Krasikova & LeBreton, 2019)，通过比较组内成员评分的实际变异与评分完全没有一致性时的期望变异之相对大小，得到误差变异的减少比例以表征组内一致性。这里的“完全没有一致性”最初被界定为“随机反应”，即所有成员的评分均匀地分布在所有选项上，由此建立的原分布称为均等分布（uniform distribution）或矩形分布。然而，评定者常带有反应偏差（response bias），使评分向某些选项发生系统性集中，此时的期望变异小于均等分布的期望变异(Bliese, 2000; James, Demaree, & Wolf, 1984; Kozlowski & Hattrup, 1992)，$r_{WG}$ 也会相应缩减。由于均等分布未能考虑反应偏差的影响，在很多时候并非刻画无一致性的最佳原分布，且计算出的结果容易高估组内一致性，已有很多学者呼吁摆脱对均等分布的过度依赖(e.g., Bliese, 2000; Brown & Hauenstein, 2005; Klein & Kozlowski, 2000)，但很难确定原分布的最佳选项。另外，$r_{WG}$ 值不够高的个别样本组是否应从后续分析中排除，也引起了一些争议。

第三，关于各指标的划界值。文献大都建议 $r_{WG}$ 和 ICC(2)以 0.7 为理想水准(e.g., Klein & Kozlowski,

2000)，但越来越多的学者提出了质疑，认为 0.7 的经验标准过于随意（arbitrary）和粗糙，只是一种缺乏理论基础的主观判断，而且将组内一致性的标准与信度的标准混为一谈根本就是错误的(Cohen, Doveh, & Nahum-Shani, 2009; Lance, Butts, & Michels, 2006; 温福星, 邱皓政, 2015)。ICC(1)常用的划界值 0.12 也有类似弊端。不过到目前为止，合适的划界值尚无定论。

聚合适当性检验可以说是多层次研究的“奠基工程”之一，发挥着“守门”和“预警”的重要作用，直接关系到高层次构念是否有合理的存在以及构念间的关系能否得到准确估计。当前，多层次研究在国内蓬勃兴起，但遗憾的是，研究者的关注点多集中于多层次模型的构建、多层次中介和调节效应分析等复杂统计方法的运用，对数据聚合中的“陷阱”和“最佳实践”则鲜有专门研究，基本上只是沿袭前人的惯用做法。本研究力图弥补这一缺憾，帮助国内学者规避概念上的误解和方法上的误用，主要目的有两个：第一，以前述三个关键问题为指引，通过对近年国内文献的系统回顾、评价以及与国外权威期刊的对照，管窥学者在聚合适当性检验中的常规实践，揭示共性问题和疏漏之处；第二，对三个关键问题进行剖析并给出实践建议。在此基础上，本研究提出应当更细致地检视和辨析各聚合指标的功能，将聚合适当性检验严格限定为组内一致性检验，各样本组的组内一致性达标后再使用组内信度指标检验构念的信度、效度，从而在“共享单位特性构念的信效度检验”的框架下将这些指标统合起来。

## 2 方法

### 2.1 期刊选择

为确保入选文献具有较高的学术水平和代表性，能够全面反映学术界对数据聚合问题的理解，我们优先从国内管理学核心期刊中选取目标期刊，入选标准为：①有公认的权威性和学术影响力。②属于国家自然科学基金委员会管理科学部认定的重要学术期刊。③发表的组织管理、组织行为方面的论文较多，以有相关专栏为宜。经讨论，最终选择了 7 份期刊，即《管理世界》《南开管理评论》《管理科学》《管理评论》《科研管理》《管理学报》《管理工程学报》。考虑到心理学期刊也发表组织行为学论文，我们又选取了《心理学报》和《心理科学》2 份期刊。

同时，为追踪国外研究现状，与国内研究形成对照，我们还选择了工业与组织心理学领域的国际权威期刊 *Journal of Applied Psychology* (*JAP*)[3]，该刊也是国外同类回顾式研究(e.g., Meyer, Mumford, Burrus, Campion, & James, 2014; Woehr et al., 2015)中常见的文献检索源刊。

[3] 作者感谢审稿专家提出的将国外代表性期刊纳入分析的建议，但由于时间和精力所限，本研究只选择了 *JAP* 一份期刊，不足以全面反映国外学者对聚合适当性问题的处理，其分析结果也不能推论至其他国外文献。不过，本研究侧重于对国内文献的回顾和评价，*JAP* 的结果主要供读者参考。

## 2.2 文献检索

从目标期刊中筛选样本文献，入选标准为：①研究中包含至少一个共享单位特性构念。②共享单位特性构念的评分来自组内个体成员评分的聚合。③明确报告了共享单位特性构念的聚合适当性检验结果（$r_{WG}$、ICC(1)、ICC(2)或方差分析结果）。④发表时间为 2014 年 1 月 1 日至 2019 年 12 月 31 日。对于中文期刊，我们以两种方式检索样本文献：一是关键词检索，在中国知网的高级检索系统中输入"聚合"、"汇聚"、"rwg"、"ICC"等关键词，并限定发表时间和期刊范围，对检索到的文献逐篇审核，确定合格文献；二是手工检索，查阅目标期刊 2014 年以来在组织行为、人力资源管理、组织管理、工商管理、创新与创业管理等栏目发表的每一篇论文，从中筛选文献。对于 *JAP*，借助 PsycInfo 数据库逐篇浏览并筛选。作者和助手先独立筛选，然后进行比对、补缺，就入选文献达成一致意见。初步入选的文献有 259 篇，为避免无效数据的干扰，我们又依据两条标准排除了 8 篇文献：①结果报告笼统，无法识别各变量检验结果的具体数值（6 篇）。②对个体层次构念进行了不必要的聚合适当性检验（2 篇）[4]。最终入选的文献共有 251 篇（中文 166 篇、英文 85 篇），详见表 1。

**表 1 样本文献发表情况**

| 期刊 | 影响因子[a] | 发表年度及数量 | | | | | | 总计 |
|---|---|---|---|---|---|---|---|---|
| | | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 | |
| 管理世界 | 7.260 | 1 | 1 | 1 | 0 | 1 | 2 | 6 |
| 南开管理评论 | 6.953 | 1 | 5 | 7 | 2 | 3 | 1 | 19 |
| 管理科学 | 5.158 | 1 | 1 | 4 | 1 | 4 | 3 | 14 |
| 管理评论 | 4.668 | 4 | 5 | 3 | 7 | 4 | 2 | 25 |
| 科研管理 | 4.280 | 4 | 4 | 2 | 2 | 5 | 5 | 22 |
| 管理学报 | 3.813 | 3 | 6 | 3 | 5 | 4 | 12 | 33 |
| 管理工程学报 | 2.968 | 0 | 3 | 5 | 3 | 1 | 1 | 13 |
| 心理学报 | 3.285 | 5 | 5 | 3 | 3 | 3 | 6 | 25 |
| 心理科学 | 1.641 | 1 | 1 | 2 | 1 | 1 | 3 | 9 |
| *JAP* | 5.067 | 23 | 21 | 16 | 9 | 11 | 5 | 85 |
| 合计 | —— | 43 | 52 | 46 | 35 | 37 | 40 | 251 |

注：[a] 中文期刊据中国知网发布的 2019 版期刊复合影响因子，*JAP* 据 JCR 2019 版影响因子。

## 2.3 编码

根据预先讨论形成的编码清单，对每一共享单位特性构念都从以下几方面进行编码：（1）构念的基本信息，包括名称、性质（自变量/因变量/中介变量/调节变量/控制变量）、题项数、计分点数、组内平均人数。（2）使用的聚合适当性检验指标及结果报告情况。（3）各指标的划界值及来源。（4）

[4] 这两篇文献均建立了低水平中介模型（2-1-1 模型或 1-1-1 模型）并计算了个体层次构念的聚合指标，我们推测这与建模时需要将层 1 变量按组均值中心化并将组均值置于层 2 截距方程式有关，但这样做是统计分析的需要（分离组间效应和组内效应），不是理论驱动下的聚合，故没有必要做聚合适当性检验。

聚合决策。

文献编码由作者和助手共同完成，具体程序为：第一，从入选的中文文献中随机抽取约 10%（15 篇）为测试样本，两人进行背靠背的独立编码，编码完毕后逐项比对，发现总编码一致性为 93.75%，分歧之处由两人讨论确定解决方案。第二，将其余文献分为两半，两人分别负责其中一半的编码工作。第三，完成各自的文献编码后，互相从对方负责的文献中随机抽取 10%（25 篇）进行二次编码和交叉复核，发现这 50 篇文献的总编码一致性为 95.40%。据此认为本研究的文献编码有较高的可信度。

# 3 结果

## 3.1 概览

166 篇中文文献中，共有 384 个变量接受了聚合适当性检验，包括自变量 142 个（36.98%）、因变量 50 个（13.02%）、中介变量 88 个（22.92%）、调节变量 93 个（24.22%）、中介+调节变量 2 个（0.52%）、控制变量 9 个（2.34%）。有 362 个变量报告了题项数量，范围在 1~42 之间，均值为 8.27（*SD*=6.45），中位数为 6，题项数不超过 6 个的变量占 54.97%，不超过 15 个的占 88.12%。有 333 个变量报告了 Likert 量表的计分点数，使用 5 点、6 点、7 点的变量分别占 55.86%、12.01%、30.63%。有 354 个变量能够识别组内平均人数，在 1.68~41.00 之间，均值为 6.02（*SD*=4.18），中位数为 5.10，人数不超过 5 的变量占 47.18%，不超过 10 的占 91.24%。使用的聚合适当性指标主要有 $r_{WG}$、ICC(1)、ICC(2)、方差分析的 *F* 检验，有 40 篇（24.10%）文献报告了 4 个指标，101 篇（60.84%）报告了 3 个指标，13 篇（7.83%）报告了 2 个指标，12 篇（7.23%）报告了 1 个指标；有 156 篇（93.98%）报告了 $r_{WG}$ 值，150 篇（90.36%）报告了 ICC(1)值，146 篇（87.95%）报告了 ICC(2)值，50 篇（30.12%）报告了 *F* 检验结果；有 137 篇（82.53%）同时报告了 $r_{WG}$、ICC(1)和 ICC(2)。聚合决策方面，有 6 个变量因 ICC(1)、ICC(2)过小或 *F* 检验不显著而被当做个体层次变量，其余变量均被聚合到高层次。

85 篇 *JAP* 文献中，共有 282 个变量接受了聚合适当性检验；题项数量在 1~23 之间（*n*=265），均值为 5.68（*SD*=3.92），中位数为 5；使用 5 点、6 点、7 点计分的变量分别占 37.59%、11.35%、37.94%（*n*=247）；组内人数在 1.63~218.99 之间，均值为 11.56（*SD*=24.49），中位数为 4.76。检验指标方面，有 71 篇（83.53%）报告了 $r_{WG}$ 值，78 篇（91.76%）报告了 ICC(1)值，73 篇（85.88%）报告了 ICC(2)值，56 篇（65.88%）报告了 *F* 检验结果[5]，61 篇（71.76%）同时报告了 $r_{WG}$、ICC(1)和 ICC(2)。有 4 个变量因为检验不达标或出于理论原因未聚合。

---

[5] *JAP* 文献中还有 3 篇使用了均差指数（average deviation index），由于该指标比较少用且与 $r_{WG}$ 相关性较强，本文暂不讨论。

3.2 $r_{WG}$ 结果报告情况

$r_{WG}$ 以单个样本组为单位计算，有学者建议，如果无法一一报告各组的 $r_{WG}$ 值，应报告所有样本组的 $r_{WG}$ 均值、中位数、范围、达到划界值的组数等汇总信息(Burke, Cohen, Doveh, & Smith-Crowe, 2018; Cohen et al., 2009; Klein & Kozlowski, 2000)。如表 2 所示，中文样本文献中有近 90%报告了 $r_{WG}$ 的均值[6]，有超过 25%报告了中位数，但 $r_{WG}$ 的范围和达到划界值的组数的报告率很低，只有 1 篇文献(张勇, 龙立荣, 贺伟, 2014)完整报告了这 4 项统计量。尤其值得注意的是，只有 3 篇文献明确报告了计算 $r_{WG}$ 值依据的原分布，其中 2 篇(韩志伟, 刘丽红, 2019; 李敏, 周恋, 2015)同时使用了均等分布和偏态分布，1 篇(邓今朝, 喻梦琴, 丁栩平, 2018)提及使用了均等分布，但都没有给出选择原分布的理由；相比之下，*JAP* 文献中有 15 篇报告了原分布，其中 1 篇同时使用了 3 种（均等分布、偏态分布和三角形分布），6 篇同时使用了 2 种（均等分布和偏态分布），8 篇使用了 1 种（均等分布和偏态分布各 4 篇），并有 5 篇给出了理由。

表 2 样本文献的 $r_{WG}$ 结果报告情况

| 报告项目 | 报告数量（按变量计）[a] | | 报告数量（按文献计）[b] | |
|---|---|---|---|---|
| | $n$ | % | $n$ | % |
| 均值 | 313[c]/157 | 87.43/72.69 | 138/53 | 88.46/74.65 |
| 中位数 | 92/76 | 25.70/35.19 | 41/29 | 26.28/40.85 |
| 范围 | 53/13 | 14.80/6.02 | 21/6 | 13.46/8.45 |
| 达到划界值的组数或比例 | 32/4 | 8.94/1.85 | 12/3 | 7.69/4.23 |
| 计算依据的原分布 | 8/31 | 2.23/14.35 | 3/15 | 1.92/21.13 |

注：表中数据，“/”左侧为中文文献，右侧为 *JAP* 文献；[a]$N_{中文}$=358，$N_{英文}$=216；[b]$N_{中文}$=156，$N_{英文}$=71；[c]有 3 个变量同时报告了均等分布和偏态分布下的 $r_{WG}$ 均值，此处不重复计数。

虽然绝大多数文献没有说明 $r_{WG}$ 值对应的原分布，但由于均等分布是研究者惯常使用的默认选项，我们参照 Woehr 等(2015)的做法，将信息缺失者均视为均等分布下的计算结果。对基于均等分布的 $r_{WG}$ 均值和中位数进行描述性统计，结果见表 3。中文文献中的变量的组内一致性总体较高，从 $r_{WG}$ 均值来看，达到 0.8 的变量超过 80%，达到 0.9 的变量超过 40%，平均值为 0.871，中位数为 0.876，只有 2 个变量的 $r_{WG}$ 均值低于 0.7，但依然进行了聚合；从 $r_{WG}$ 中位数来看，达到 0.9 的变量占 70%，平均值为 0.908。*JAP* 文献中的 $r_{WG}$ 值的各项统计指标均略低于中文文献。另外，中文文献中使用偏态分布计算的 3 个变量的 $r_{WG}$ 均值分别为 0.84、0.93、0.70，*JAP* 文献中使用偏态分布计算的 12 个 $r_{WG}$ 均值在 0.67~0.97 之间。

[6] 有 49 篇文献（中文 31 篇、英文 18 篇）没有说明报告的是 $r_{WG}$ 的均值还是中位数，大多模糊地称为“$r_{WG}$ 值”，考虑到均值是最常用的统计量，故都按均值对待。

表 3 样本文献中基于均等分布的 $r_{WG}$ 值的描述性统计

| 来源 | 统计量 | *n* | *M* | *SD* | *Me* | 范围 | 达到相应值的变量数量及比例 | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | ≥0.7 | ≥0.8 | ≥0.9 |
| 中文文献 | $r_{WG}$ 均值 | 313 | 0.871 | 0.071 | 0.876 | 0.572~0.990 | 311 (99.36%) | 265 (84.66%) | 134 (42.81%) |
| | $r_{WG}$ 中位数 | 92 | 0.908 | 0.067 | 0.926 | 0.750~0.980 | 92 (100%) | 84 (91.30%) | 65 (70.65%) |
| *JAP* 文献 | $r_{WG}$ 均值 | 148 | 0.840 | 0.084 | 0.840 | 0.630~0.990 | 142 (95.95%) | 102 (68.92%) | 42 (28.38%) |
| | $r_{WG}$ 中位数 | 74 | 0.878 | 0.089 | 0.895 | 0.610~0.990 | 70 (94.59%) | 61 (82.43%) | 37 (50.00%) |

注：*Me*=中位数。

### 3.3 ICC(1)结果报告情况

由表 4 可知，中文文献中 ICC(1)值的均值为 0.276，中位数为 0.250，90%达到了最常被引用的划界值 0.12。有 19 个变量的 ICC(1)值低于 0.1，其中 4 个未聚合。另有 114 个变量报告了方差分析的 *F* 检验结果，只有 1 个变量未达到 0.05 的显著水平，作者也做出了不聚合的决定。

*JAP* 文献中 ICC(1)值的均值和中位数分别为 0.241、0.210，达到 0.12 的比例亦低于中文文献；有 32 个变量的值低于 0.1，其中 3 个未聚合。163 个 *F* 检验结果中有 3 个不显著，其中 1 个未聚合。

表 4 样本文献中 ICC(1)值的描述性统计

| 来源 | *n* | *M* | *SD* | *Me* | 范围 | 达到相应值的数量及比例 | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | ≥0.12 | ≥0.20 | ≥0.30 | ≥0.40 |
| 中文文献 | 336[a] | 0.276 | 0.141 | 0.250 | 0.011~0.790 | 304 (90.48%) | 231 (68.75%) | 127 (37.80%) | 61 (18.15%) |
| *JAP* 文献 | 247 | 0.241 | 0.157 | 0.210 | 0.010~0.851 | 195 (78.95%) | 132 (53.44%) | 69 (27.94%) | 39 (15.79%) |

注：*Me*=中位数；[a] 不包括只笼统地报告了 ICC(1)>0.05 的 3 个变量。

### 3.4 ICC(2)结果报告情况

由表 5 可知，中文文献中 ICC(2)值的均值为 0.695，中位数为 0.714，达到传统划界值 0.7 的刚刚超过一半；有 38 个变量的 ICC(2)值低于 0.5，其中 5 个未聚合。*JAP* 文献中 ICC(2)的均值和中位数分别为 0.596、0.630，达到 0.7 的仅有三分之一；有 70 个变量的值低于 0.5，其中 4 个未聚合。

表 5 样本文献中 ICC(2)值的描述性统计

| 来源 | *n* | *M* | *SD* | *Me* | 范围 | 达到相应值的数量及比例 | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | ≥0.6 | ≥0.7 | ≥0.8 | ≥0.9 |
| 中文文献 | 322[a] | 0.695 | 0.177 | 0.714 | 0.047~0.991 | 241 (74.84%) | 185 (57.45%) | 93 (28.88%) | 35 (10.87%) |
| *JAP* 文献 | 232 | 0.596 | 0.204 | 0.630 | 0.100~0.980 | 133 (57.33%) | 84 (36.21%) | 36 (15.52%) | 14 (6.03%) |

注：*Me*=中位数；[a] 不包括只笼统地报告了 ICC(2)>0.5 的 3 个变量。

3.5 划界值引用情况

有 112 篇（67.47%）中文文献给出了至少一个聚合指标的划界值，64 篇（38.55%）引用了来源文献；*JAP* 文献中有 20 篇（23.53%）给出了划界值，44 篇（51.76%）引用了文献，见表 6。几乎所有 $r_{WG}$ 值都使用了 0.7 的标准，ICC(1)的划界值以 0.12 和 0.05 为多，ICC(2)最常用的为 0.5 和 0.7（英文文献亦有多篇使用 0.6）。中英文文献在划界值的源文献上有较大分歧：在中文文献中，James 的 3 篇经典文献(James, 1982; James et al., 1984; James, Demaree, & Wolf, 1993)的被引次数遥遥领先，一项关于服务氛围和服务质量的实证研究(Schneider, White, & Paul, 1998)也得到了多次引用（ICC(2)的一个"划界值"0.47 出自该文）；*JAP* 文献最常引用的则是 Bliese (1998, 2000)与 LeBreton 和 Senter (2008)，提及 James 和 Schneider 文的明显较少。

表 6 样本文献使用的划界值统计

| 指标 | 划界值 | 被引次数 | 主要源文献 |
|---|---|---|---|
| $r_{WG}$ | 0.7 | 96/12 | James 等(1982, 1984, 1993)（41/10）[a] |
| ICC(1) | 0.05 | 28/1 | Bliese (1998, 2000)（12/28） |
| | 0.1 | 11/0 | LeBreton & Senter (2008)（2/18） |
| | 0.12 | 32/4 | Schneider 等(1998)（10/1） |
| ICC(2) | 0.47 | 10/0 | Klein & Kozlowski (2000)（4/3） |
| | 0.5 | 34/0 | Glick (1985)（0/5） |
| | 0.6 | 7/4 | 张志学(2010)（3/0） |
| | 0.7 | 30/7 | |

注：表中数据，"/"左侧为中文文献，右侧为 *JAP* 文献；源文献与左侧的划界值无对应关系；[a] 文献后括号里的数字表示被引次数。

# 4 讨论

对国内 9 份管理学、心理学期刊 2014 年以来的文献分析表明，绝大多数包含共享单位特性构念的多层次研究都将聚合适当性检验视为数据分析的前置程序，广泛使用 $r_{WG}$、ICC(1)、ICC(2)等指标为聚合提供实证证据。从各指标的汇总结果来看，$r_{WG}$ 和 ICC(1)值普遍较高，ICC(2)相对略低。Woehr 等(2015)回顾了 1998~2012 年发表于 *Academy of Management Journal* 等 4 份权威期刊的 189 篇文献，发现基于均等分布的 $r_{WG}$ 平均值为 0.84（$n$=486），约有 90%的值高于 0.7，近 30%的值高于 0.9；ICC(1)均值为 0.21（$n$=416），超过 75%的值高于 0.11；ICC(2)均值为 0.66（$n$=372），只有近一半的值高于 0.7。本研究还汇总了 *JAP* 的文献，发现 $r_{WG}$、ICC(1)和 ICC(2)的平均水平分别是 0.840、0.241、0.596。总体上，与 *JAP* 文献相比，国内文献中 $r_{WG}$、ICC(1)和 ICC(2)的平均水平（0.871、0.276、0.695）更高，达到划界值的比例也略胜一筹，这从一个侧面揭示，国内优秀期刊发表的文献的数据质量已达到国际主流水平。

另一方面，国内研究在聚合适当性检验中也存在一些薄弱环节，以前述三大关键问题的分析视角，可以归结为下列三点：

第一，对组内一致性和组内信度的功能未加区分。大多数研究将 $r_{WG}$、ICC(1)、ICC(2)视为表征聚合适当性的平行指标，只关心计算结果是否达到了“门槛”，对其理论意涵和独特用途思考不多，一旦组内一致性和组内信度的结果出现矛盾（如 $r_{WG}$ 值很高但 ICC 值较低），在解释结果和做出聚合决策时就会面临两难境地；还有个别研究将组间差异与组内成员的共识程度视为等价，“绕过”组内一致性而仅依据 ICC 值做出聚合决策，有构念误设（construct misspecification）的风险。这一问题在 *JAP* 文献中也较普遍。

第二，计算 $r_{WG}$ 时未能审慎选择原分布。国内研究者普遍将均等分布视为计算 $r_{WG}$ 时“缺省”甚至唯一的原分布，甚至认为没有必要赘述这一“不言自明”的常识，结果就是样本文献中仅有 3 篇明确报告了原分布，仅有 2 篇使用了均等分布以外的原分布。作为对比，Meyer 等(2014)检视了 111 篇英文文献中的 440 个 $r_{WG}$ 值，发现 24.1%的值报告了原分布，其中只有 69.8%使用的是均等分布；在 Woehr 等(2015)的文献回顾中，有近 10%的 $r_{WG}$ 值使用了轻度偏态分布；本研究编码的 *JAP* 文献中有 21.1%（15/71）报告了原分布，其中 73.3%（11/15）使用了偏态分布。可见，虽然国外研究也有默认使用均等分布的“通病”，但原分布的报告率和替代性原分布的使用率都明显高于国内研究。另一个问题是结果报告不够完整。大部分文献只报告了所有组的 $r_{WG}$ 均值或中位数，忽略了 $r_{WG}$ 值的范围和达到划界值的组数。$r_{WG}$ 均值和中位数只是对所有样本组的 $r_{WG}$ 值集中趋势的刻画，不足以体现 $r_{WG}$ 值在组间的离散和分布情况，因为较高的 $r_{WG}$ 均值并不代表所有组的 $r_{WG}$ 值都能达标，完全可能存在个别 $r_{WG}$ 值很低、达不到聚合基本要求的组。这类无效样本组只能通过逐一检视各组的 $r_{WG}$ 值来识别。

第三，划界值选取杂乱，部分引用有误。由于学界对各聚合指标的划界值尚未达成共识，稳妥的做法是根据研究问题和情境预先设定好划界值(Biemann, Cole, & Voelpel, 2012)并准确引用源文献。中文样本文献中虽然有三分之二指出了选用的划界值，但只有约 40%引用了源文献，而且如果细加查证，就会发现不少引用是错误的。例如，很多研究在介绍 $r_{WG}$、ICC(1)和 ICC(2)的划界值时只引用了 James 的某一篇文献（如最著名的 1984 年发表于 *JAP* 的论文），实际上除了能在 James (1982)中找到一个日后被众多学者“误读”的所谓 ICC(1)的“标准”（0.12）[7]外，这几篇文献没有推荐或提及任何一个指标的划界值；ICC(2)的划界值更是出现了 0.47、0.5、0.6、0.7 等多个，其中 0.47 是 Schneider 等(1998)

[7] James (1982, p.224)回顾了 1970 年代对组织气氛的多项研究，发现组内一致性的中位数为 0.12，但这不单纯是 ICC(1)的结果，还包含了 $\eta^2$ 和 $\omega^2$（这几个指标当时被认为反映了组内一致性），因此 0.12 既不是明确提出的划界值，更不是 ICC(1)的合法划界值。

从自己的研究数据中算得的，0.5 的来源则无据可考，也许只是 0.47 的近似值，把这两个值当成划界值显然不合适。这些疏漏恐怕是由于研究者没有仔细查证原文就照搬前人的做法，导致以讹传讹。相比之下，*JAP* 的源文献引用率更高、引证更准确，使用的也大都是“正统”的经验标准。

## 5 关键问题试解与实践建议

从国内研究暴露出的普遍性问题来看，引言中提出的三个未解难题已成为正确、规范运用聚合适当性检验的障碍，但我们认为这些难题之所以会引起疑惑，不在于统计原理的高深复杂，而在于研究者对基本理论缺乏明察和深究。为此，本部分将从理论和经典文献出发，对这些难题进行逐一剖析，并尝试提出便于应用型研究者掌握的实践建议。

### 5.1 聚合指标的选择

学界的基本共识是，组内一致性和组内信度服务于不同的研究目的，并非相互替代或竞争的关系，而是从不同角度提供了关于共享单位特性构念信效度的信息，在聚合检验中应同时使用。但是，国内学者往往将二者的地位和功能简单等同起来并不加区别地使用，为结果的解释带来困惑(徐晓锋, 刘勇, 2007)。以下将从理论和实证两方面进行讨论和澄清。

首先，如果深入到对共享单位特性构念的理论思考中，就会发现组内一致性和组内信度扮演着不同的角色。共享单位特性构念的理论意义存在于团体层次，从心理测量学的角度看就是只有组水平的真分数(Newman & Sin, 2020)，但该分数的源头是团体内个体成员的态度、感知、价值观等，并经由社会化、领导、内部互动等心理过程的影响逐步形成共同的认知(Kozlowski & Klein, 2000)；如果不具备这种共同认知，或者说成员的态度、想法不一致，共享单位特性构念就失去了存在的根基，也就是团体内无法形成一个能够有效代表成员共同认知的集合构念(Moritz & Watson, 1998; 于海波等, 2004)，个体数据的聚合也没有意义(方杰, 张敏强, 邱皓政, 2010)。可见，聚合的首要标准是看团体成员的意见一致性/共享性（sharedness）如何，这只能由组内一致性指标“捕捉”到，即聚合适当性检验的实质是组内一致性检验，组内一致性达标表示组内成员评分的均值是共享单位特性构念的适宜代理值(Cohen et al., 2001; Dunlap, Burke, & Smith-Crowe, 2003; Van Mierlo, Vermunt, & Rutte, 2009)，可以将个体评分聚合到单位层次。

相比之下，组内信度不直接指向组内成员的意见一致性，而是关心组间差异性或区分度(Chan, 1998; Kirkman, Tesluk, & Rosen, 2001; Quigley, Tekleab, & Tesluk, 2007; Van Mierlo et al., 2009)，其之所以得到极大关注，或许根本原因是它触及研究者的一个重大关切：缺乏组间变异可能导致统计检验力降低、II 型错误率上升，使该构念对其他构念的预测力被低估(Bliese, 1998; Bliese, Maltarich,

Hendricks, Hofmann, & Adler, 2019; George & James, 1993; James, 1982; Moritz & Watson, 1998)，削弱研究结果的可信度。从这个角度看，组内信度检验更像是为聚合后的共享单位特性构念加了一道“保险”：确保该构念可以与其他构念产生有意义的关联(James, 1982)，确保路径系数估计值准确无偏。

其次，从实证上看，组内信度只能在一定程度上间接推断组内一致性，且不够准确。考察发展脉络可知，数据聚合问题衍生自对组织气氛（organizational climate）的研究，最初用来衡量聚合适当性的指标其实是 ICC(1)(James, 1982)，因为 ICC(1)的计算实质是比较组间变异与组内变异的相对大小，ICC(1)较大意味着组间变异较大、组内变异较小，也就是组内一致性很高、随机性很低(Bartko, 1976; 罗胜强, 姜嬿, 2014)。但问题在于，组内变异和组间变异并非此消彼长的关系，二者可以同时都很大，此时就会出现 ICC 很高但组内一致性很低的矛盾现象(Bliese, 2000; Kozlowski & Hattrup, 1992)。例如，表 7 呈现了 6 个 4 人小组对某一题项的评分，由组均分的差异可以推测组间变异较大，计算结果为 ICC(1)=0.74，ICC(2)=0.92；但也容易观察到第 5、6 组成员的评分有不小的组内分歧，进一步算得这两组的 $r_{WG(1)}$值分别只有 0.26、0.45。这清楚地揭示出组内信度的严重缺陷，即不能提供各组组内一致性的详细信息，而且对一致性不足的组不敏感。我们在分析中也发现样本文献中的 ICC(1)、ICC(2)与 $r_{WG}$ 均值的相关均不显著（$r$=0.08、–0.06，$ps$>0.05），进一步表明组内信度与组内一致性并无实质联系。

由此，组内信度在理论上不直接触及组内一致性，在统计上不能准确估计组内一致性，不适合作为聚合适当性的指标。但也要看到，ICC(1)和 ICC(2)在共享单位特性构念的信度、效度检验上发挥着很大作用。ICC(1)反映组间差异性，它不是共享单位特性构念存在的核心证据，而是检测构念之间关系的要素(廖卉, 庄瑷嘉, 2012)，更适合作为共享单位特性构念的效度指标。ICC(2)是组内所有成员评分的均值的可靠性，即组均值的信度(Bliese, 2000; Dixon & Cunningham, 2006)，全样本的 ICC(2)实际上是所有组的组均值信度的平均水平。若以经典测量理论来解释，ICC(2)类似于权重，衡量了样本均值在组真值（未知参数）估计中的贡献度，组间变异越大、组内人数越多，由样本计算的组均值的可靠性越高，越能代表组真值的潜在水平；如果组均值的信度不高，就必须更多地“借力”于总体均值的估计值来推测组真值(Raudenbush & Bryk, 2002; 温福星, 邱皓政, 2015)。可见，ICC(2)体现的是用组内成员评分的均值作为该组在某构念上的潜在水平的信心程度。由于最常用的信度指标 Cronbach’s α 需要以所有评定者来自独立同质的总体和单层次因子结构为前提，不适用于违反非独立性假设的嵌套数据(Geldhof, Preacher, & Zyphur, 2014; 温福星, 邱皓政, 2015)，建议以多层次数据的专用指标 ICC(2)取代单层次 Cronbach’s α 做共享单位特性构念的信度指标(e.g., Jiang, Chuang, & Chiao, 2015)，

注意基于 ICC(2)的信度表示的是组均值的代表性而不是题项的内部一致性。

综合以上分析，组内一致性和组内信度的新功能定位如下：组内一致性最符合共享单位特性构念的理论规定，是聚合适当性的最重要标准，回答的是“组内成员的意见是否足够一致”或“共享单位特性构念是否有合理的存在”的问题；ICC(1)是效度指标，回答的是“构念在组间的差异是否足够”或“构念间的关系能否被准确估计”的问题；ICC(2)是信度指标，回答的是“组均值对组真值的代表性是否足够”的问题。组内一致性和组内信度都属于共享单位特性构念之信效度检验的必要程序，但聚合决策应主要依据组内一致性信息。实践中的理想情形是组内一致性和组内信度都较高，这时做出聚合决策就有充分的信心，但也可能遇到以下两种矛盾情形：

a.各组的 $r_{WG}$ 值都较高，但全样本的 ICC 值较低。

b.全样本的 ICC 值较高，但个别组的 $r_{WG}$ 值较低。

第一种情形暗示各组均值非常接近（应留意得分是否集中于量表的高分端或低分端），虽然在理论上可以聚合，但要承担统计检验力降低和构念间关系的估计值出现偏误的风险，如果 ICC(1)和 ICC(2)都很低，数据分析结果就不可信；在第二种情形下，组内一致性不达标的组的得分不够稳定，会为整体得分引入随机误差，损害构念的效度，建议将其剔除（详见下一节）。

表 7  示例数据

| | 评分者 A | 评分者 B | 评分者 C | 评分者 D | 组均分 | 组内变异 | $r_{WG(1)}$[a] |
|---|---|---|---|---|---|---|---|
| 小组 1 | 2 | 4 | 3 | 3 | 3.00 | 0.67 | 0.90 |
| 小组 2 | 5 | 7 | 5 | 6 | 5.75 | 0.92 | 0.86 |
| 小组 3 | 1 | 3 | 1 | 2 | 1.75 | 0.92 | 0.86 |
| 小组 4 | 7 | 9 | 9 | 8 | 8.25 | 0.92 | 0.86 |
| 小组 5 | 2 | 4 | 6 | 1 | 3.25 | 4.92 | 0.26 |
| 小组 6 | 6 | 8 | 8 | 4 | 6.50 | 3.67 | 0.45 |

资料来源：罗胜强和姜嬿(2014)，p.276；表中内容有增删。

注：[a] 基于 9 点量表和均等分布计算的结果。

### 5.2 $r_{WG}$ 的计算与数据清理

$r_{WG}$ 计算中的最大困难是选择合适的原分布，对此尚无完美的解决方案。作为权宜之计，我们建议研究者响应国外学者的呼吁(Biemann et al., 2012; Castro, 2002; James et al., 1984; Kozlowski & Hattrup, 1992; LeBreton & Senter, 2008)，不要仅使用最保守、结果最“理想”也可能最不准确的均等分布，而应将均等分布下的计算结果视为组内一致性的近似上界，另参考理论和已有研究选择一种替代性原分布来估计组内一致性的近似下界，这样可大致确定组内一致性真值的范围。在选择替代性原分布时，应仔细考量评定者可能存在的反应偏差。在组织管理情境中一些非随机因素的作用下，成员

的反应容易带有社会赞许性、趋中偏差（central tendency bias）、宽大偏差（leniency bias）/严苛偏差（severity bias）等，它们会导致“随机反应”偏离均等分布，呈现三角形分布或偏态分布(James et al., 1984; Klein et al., 2001; Ng, Koh, Ang, Kennedy, & Chan, 2011; Smith-Crowe, Burke, Cohen, & Doveh, 2014; Smith-Crowe, Burke, Kouchaki, & Signal, 2013)；另外，组织中的某些社会、心理、政治因素会成为强情境线索，对成员的反应偏差造成普遍的系统性影响(Meyer et al., 2014)。表 8 总结了组织中常见的反应偏差及相应的原分布，研究者可参照选择替代性原分布并在文中给出具体的理由。一般而言，在评价同事、主管和自己的团队时，评定者倾向于给出比实际情况更加积极的评价，因此轻度偏态分布是有广泛适用性的推荐选项(e.g., Rego, Cunha, & Simpson, 2018; Schaubroeck, Shen, & Chong, 2017)。

$r_{WG}$ 计算完成后，研究者可能会发现有一些样本组未能达到合格标准，接下来要决定是否剔除这些组，学界对此立场不一：有的学者较为宽容，主张只要全样本的 $r_{WG}$ 均值或中位数达标，就可将所有组纳入后续分析，无需剔除不合格的组，否则会导致样本量减少、统计检验力降低、II 型错误率上升(Carron et al., 2003; LeBreton & Senter, 2008)；也有学者建议进行敏感性分析，即在剔除和不剔除两种情况下分别分析数据，比较结果是否有显著差异(Biemann et al., 2012; Woehr et al., 2015)；还有学者明确指出不合格的组不应保留，否则会导致构念间的效应缺失（missed）、虚假效应（misidentified）或错误解释（misinterpreted）(Castro, 2002; Van Mierlo et al., 2009)。从实践来看，国外大多数研究都选择保留所有样本组(Burke et al., 2018)，我们也发现样本文献中仅有 3 篇(Farmer, Van Dyne, & Kamdar, 2015; 吕洁，张钢, 2015; 马君，张昊民，杨涛, 2015)明确报告删除了 $r_{WG}$ 值未达标的组。究竟是否应当排除不合格的组？其实这不是统计问题，而是理论问题，只要回到共享单位特性构念的本质内涵上就不难找到答案。前面已论述过，组内成员的共同认知才是共享单位特性构念存在的根基，组内一致性很差提示团队成员不够团结(Moritz & Watson, 1998)，甚至已分裂为亚组（subgroup）或“小帮派”(Castro, 2002; LeBreton, James, & Lindell, 2005)，根本无法形成共识，强行聚合不但违反了共享单位特性构念的理论假设，而且不可靠的组均值会为构念的测量引入误差。因此，建议研究者坚守严格的标准，将组内一致性不达标的组排除[8]，对剩余样本组再次检验 ICC(1)和 ICC(2)以确认构念的信效度。

[8] 虽然不满足聚合的条件，较低的组内一致性仍有两方面研究价值：其一，在离散模型（dispersion model）中，作为表征组内差异性的独立构念（如气氛强度和领导—成员交换差异化）；其二，在共识涌现模型（consensus emergence model）中，揭示组织成员在价值观、信念、行为等方面的一致性的动态发展过程。感兴趣的读者可查阅相关文献(e.g., Lang, Bliese, & de Voogt, 2018; Lang, Bliese, & Runge, in press; 蒋丽，李永娟，田晓明, 2012)。

表 8 组织管理研究中常见的反应偏差及原分布

| 反应偏差 | 描述 | 对应的原分布 | 部分适用情境 |
|---|---|---|---|
| 社会赞许性 | 指评定者倾向于按照他人期望的方式做出评定 | 偏态分布（轻/中/重度偏态） | 1.测量的是对工作环境的感知和评价（特别是带有负效价的构念），如团队负性情绪氛围、团队冲突、辱虐管理等<br>2.无法保证问卷的匿名性（例如在领导—部属配对调查中使用编号、代号等以便识别评定者的身份） |
| 趋中偏差 | 指评定者倾向于隐藏真实态度，选择中立的选项 | 三角形分布<br>正态分布 | 1.题项含义模糊、表述不清或过于复杂<br>2.评定者缺乏专门培训<br>3.评定者缺乏参与动机，不愿表明态度<br>4.做出的回答牵涉到评定者的个人利益且无法保证匿名<br>5.集体主义文化中的个体评价自己的绩效和工作表现 |
| 宽大偏差 | 指评定者倾向于做出比自己的真实态度更加积极的评定 | 偏态分布（轻/中/重度偏态） | 1.评价主管的积极领导力和其他组织所重视的优良特质时（在进行面对面或非匿名的评价时，或评定者具有高权力距离取向和集体主义价值观时，宽大偏差会加重，导致中度到重度偏态）<br>2.评价同事和团队的绩效和其他积极特质时<br>3.主管为了得到部属的支持或展现自己的领导能力，在评价部属的绩效时会打分偏高 |

最后还有两点提示：第一，研究者大多希望 $r_{WG}$ 值越高越好，但单个组过高的 $r_{WG}$ 值（如高于 0.97）也是一个警示信号，暗示成员的评分可能多集中于量尺的端点（最高分或最低分）(Carron et al., 2003)。此时应检查原始数据，如果情况属实，不排除有外力介入（如主管的诱导、指示）或无效施测（如相互传抄或指定某人代填）的可能，特别是 $r_{WG}$ 值为 1 的组[9]嫌疑更大，建议将这种呈现“可疑一致性”的组当做异常值剔除。当然，准确鉴别合理的高一致性和可疑一致性有赖于研究者的经验，但更重要的是在量表编制和调查实施阶段做好质量控制。第二，关于结果报告时 $r_{WG}$ 均值和中位数的选择，建议研究者同时报告这两个值，理想情况下它们应当非常接近，但如果二者相差较大，提示可能存在极端组，有必要逐一检查，寻找组内一致性过低或过高的组。

## 5.3 划界值的选取

组内一致性和组内信度的合格标准是长久以来的争论焦点。$r_{WG}$ 和 ICC(2)最广为接受的划界值之所以是 0.7，是因为早期文献将这两个指标都归为信度的范畴，虽然 $r_{WG}$ 后来被修正为组内一致性的指标，但 0.7 的划界值却沿用至今；ICC(1)最常用的划界值 0.12 仅源于 James (1982)对少量文献的结

[9] 这表明组内成员的所有评分完全相同，组内变异为 0，因此无论使用哪种虚无分布计算，$r_{WG}$ 始终等于 1。

果汇总。这些武断的经验标准缺乏坚实的理论根基，而且存在简单化倾向，未能精细地考虑组内人数、题项数量、计分点数等因素的潜在影响(Cohen et al., 2001; Lance et al., 2006; LeBreton & Senter, 2008)，遭到了越来越多的批评。

为摆脱对经验标准的依赖，部分学者寻求对组内一致性进行显著性检验，找到客观且有统计学依据的划界值，“另起炉灶”建立一套统计标准。基本思路是，预先设定若干背景条件（如组内人数、题项数量、计分点数、题项间的平均相关系数、原分布），再使用基于 Monte Carlo 模拟的近似随机化检验（approximate randomization test）或随机组重取样法（random group resampling）生成海量模拟数据，找出各种条件组合下 $r_{WG}$ 值的 95%百分位数作为临界值(Cohen et al., 2001; Cohen et al., 2009; Dunlap et al., 2003; Smith-Crowe et al., 2014)[10]。从假设检验的角度看，其目的是推断样本来自的总体是仅具有巧合或偶然的组内一致性（chance agreement），还是具有系统的组内一致性(Dunlap et al., 2003; O'Neill, 2017)。统计标准克服了经验标准的主观性弊端，但也有两个突出问题：第一，设定的条件只是一些典型值，远无法涵盖实际研究中的所有情况，常常难以找到与研究的具体条件完全契合的精确临界值。第二，达到统计显著性只是拒绝了“不存在组内一致性”的虚无假设，但不能保证组内一致性足够高。表 9 展示了部分条件组合下 $r_{WG}$ 达到 0.05 的显著性水平时的临界值，可知组内人数为 5 人时临界值较高（在 0.8 左右），而组内人数达到 10 人且题项较少时，临界值明显降低，甚至低于 0.7 的经验标准，这样即使在统计上显著，实际意义也不大。受此限制，统计标准不能很好地满足研究需要。

就达到聚合所需的充分的组内一致性而言，经验标准有更高的实用价值(Lüdtke & Robitzsch, 2009; O'Neill, 2017)，仍可作为聚合决策的主要依据，但需要进行修正和改进，目前有两条路径：一是将“通过—不通过”的二分式评判细化为类似效应量评价的等级制，如区分为小效应、中效应、大效应；二是以现有研究的平均水平为参照系，如 Woehr 等(2015)对近 200 篇文献的汇总结果。我们力图将这两种策略加以整合，尝试性地提出组内一致性和组内信度的新标准。具体而言，对于均等分布下的 $r_{WG}$，Woehr 等从文献中汇总的 $r_{WG}$ 均值为 0.84，本研究汇总的结果为 0.87/0.84（中文文献/*JAP* 文献），而 Brown 和 Hauenstein (2005)、LeBreton 和 Senter (2008)划定的“强一致性”的标准分别是 $r_{WG}$≥0.8、0.71≤$r_{WG}$≤0.90，故建议 $r_{WG}$ 的临界值在均等分布下设为 0.8，在轻度偏态分布下稍微放宽，设为 0.7。对于 ICC(1)，LeBreton 和 Senter (2008)提出 0.01、0.1、0.25 可分别对应于小效应、中效应、大效应，本研究和 Woehr 等汇总的平均值分别为 0.276/0.241（中/*JAP*）、0.21，我们建议以达到 0.2 为佳，或

---

10 本研究选取的英文文献中有 4 篇对 $r_{WG}$ 值进行了这种显著性检验。

要求方差分析的 $F$ 检验至少达到 0.01 的显著性水平，确保有较大的效应量。ICC(2)是 ICC(1)和组内人数的函数，评分者的增加会使 ICC(2)随之提高，但组织管理研究的组内人数通常不多[本研究和 Woehr 等的汇总结果分别为 5.10/4.76（中/*JAP*，中位数）、6.93]，取得较高的 ICC(2)相对困难，0.7 的常规标准略显严苛（本研究的中、英文样本文献分别只有一半和三分之一的 ICC(2)超过了 0.7）。考虑到 Glick (1985)曾在讨论组织气氛的测量问题时提出，无论采用哪种计算指标，聚合后的组均值的信度需要达到 0.6(也见 杨建锋, 王重鸣, 2008)，该标准在 *JAP* 中也多次被引用，我们认为在 ICC(1)达标的前提下，可以把 0.6 作为 ICC(2)可接受的下限，建议平均组内人数少于 8 人时放宽标准至 0.6，达到 8 人时取 0.7，如果人数过多（如超过 20）最好进一步提高标准以抑制 ICC(2)的膨胀效应。

当然，为基本条件千差万别的研究设定统一的划界值并不“公平”，因此不宜固守一成不变的标准，而应容许适度的变通空间(Krasikova & LeBreton, 2019)。研究者可以根据实际情况采用稍加严格或宽松的标准，但应有理有据，并在分析数据之前就设定好，不可随意更改。最重要的是，研究者要加强理论思考，克服将经验标准绝对化的不良倾向，避免“把研究的责任交给计算机和这些机械的判定标准”(辛自强, 2018, p.346)。

**表 9 部分条件组合下 $r_{WG(J)}$的临界值**

| 原分布 | $N$ | ρ | 5 点计分 | | | 7 点计分 | | |
|---|---|---|---|---|---|---|---|---|
| | | | 3 题 | 5 题 | 10 题 | 3 题 | 5 题 | 10 题 |
| 均等 | 5 | 0.4 | 0.74 | 0.76 | 0.83 | 0.75 | 0.77 | 0.84 |
| | 5 | 0.6 | 0.79 | 0.81 | 0.88 | 0.78 | 0.82 | 0.88 |
| | 10 | 0.4 | 0.57 | 0.63 | 0.73 | 0.57 | 0.64 | 0.74 |
| | 10 | 0.6 | 0.61 | 0.68 | 0.78 | 0.61 | 0.70 | 0.80 |
| 三角形 | 5 | 0.4 | 0.78 | 0.81 | 0.86 | 0.78 | 0.81 | 0.86 |
| | 5 | 0.6 | 0.82 | 0.85 | 0.90 | 0.82 | 0.85 | 0.90 |
| | 10 | 0.4 | 0.65 | 0.70 | 0.78 | 0.64 | 0.70 | 0.79 |
| | 10 | 0.6 | 0.68 | 0.75 | 0.84 | 0.69 | 0.76 | 0.84 |
| 轻度偏态 | 5 | 0.4 | 0.80 | 0.83 | 0.88 | 0.81 | 0.83 | 0.88 |
| | 5 | 0.6 | 0.85 | 0.87 | 0.91 | 0.84 | 0.87 | 0.92 |
| | 10 | 0.4 | 0.67 | 0.73 | 0.80 | 0.67 | 0.72 | 0.81 |
| | 10 | 0.6 | 0.71 | 0.77 | 0.85 | 0.71 | 0.77 | 0.85 |

资料来源：根据 Smith-Crowe 等(2014)的模拟研究结果整理而成。

注：$N$=组内人数；ρ=题项间的平均相关系数；实际计算的 $r_{WG(J)}$值大于表中的临界值表示具有统计显著性($p$<0.05)。

最后，整合上述实践建议，本研究提出一套包含聚合适当性检验在内的共享单位特性构念的信效度检验程序（表 10），研究者可参照执行，并将检验结果呈现于论文“研究结果”部分的第一节（附录提供了各指标的计算工具）。还需说明两点：（1）按照逻辑顺序，应当先检验聚合适当性，再检验效度和信度，因为只有通过了聚合适当性检验，才能确认高层次构念的聚合分数有效，可用于后续

分析。（2）在效度检验环节，不少研究会把高层次构念与其他个体层次构念放在一起进行验证性因子分析以检验区分效度，但这种做法忽视了数据的嵌套特性，将高层次构念“降级”为低层次构念，混淆了组内和组间因子结构，是错误的。正确的做法是对高层次构念单独执行多层次验证性因子分析，同时分析组内和组间协方差矩阵，重点考察模型的整体拟合度以及组间结构的因子负荷是否理想(Dyer, Hanges, & Hall, 2005; 王孟成, 毕向阳, 2018)；还可进一步计算组间的 Cronbach’s α 和组合信度（ω 系数、*H* 系数）等(Geldhof et al., 2014; 田雪垠, 郑蝉金, 郭少阳, 贺冠瑞, 2019)，达成心理测量学信度（psychometric reliability）与聚合信度（aggregate reliability，即 ICC(2)）的互补(Jebb, Tay, Ng, & Woo, 2019)。

**表 10 共享单位特性构念的信效度检验程序**

| 步骤 | 推荐做法 | 不恰当的做法 | 补充说明 |
|---|---|---|---|
| I.准备阶段 | 1.阐明共享单位特性构念所在的层次及理论依据<br>2.报告预先设定的各指标的划界值，需简述理由或引用相关文献 | 1.忽视对构念所在层次的思考和讨论<br>2.固守陈旧的不合理的划界值；无根据地任意选取划界值；不明确报告划界值；文献引用不当 | 对构念所在层次的阐述应置于理论模型构建部分（先于方法部分） |
| II.聚合适当性检验 | 1.报告拟使用的原分布（至少 2 种）及理由<br>2.分别报告各原分布下全样本的 $r_{WG}$ 均值、中位数、范围、达到划界值的组的比例、因不达标而被剔除的组的数量 | 1.只使用均等分布<br>2.无根据地选取原分布<br>3.结果报告不全，如只报告 $r_{WG}$ 均值<br>4.对 $r_{WG}$ 值不达标的组不加处理 | 应检查每个样本组的 $r_{WG}$ 值，将 $r_{WG}$ 值不达标和过高的组排除出后续分析；如果不合格的组过多，建议检查施测过程、补充数据 |
| III.效度检验 | 1.报告 ICC(1)值和方差分析的 *F* 检验结果<br>2.报告多层次验证性因子分析结果 | 把高层次构念与个体层次构念放在一起，执行常规的单层次验证性因子分析 | 1.如果 ICC(1)较小，构念间的关系可能被低估，应在文中讨论这种局限性<br>2.如果组数较少，多层次验证性因子分析容易出现收敛困难或估计偏差 |
| IV.信度检验 | 1.报告 ICC(2)值（聚合信度）<br>2.如可以实现多层次验证性因子分析，还应计算组间的 Cronbach’s α 或 ω 系数、*H* 系数（心理测量学信度） | 忽视组间结构，以单层次的 Cronbach’s α 作为整体信度指标 | 当组内人数较多时，ICC(2)容易膨胀，需确保 ICC(1)足够大；当组内人数较少时，如果 ICC(1)较大，略小的 ICC(2)亦可接受 |

## 6 结语

围绕多层次研究的数据聚合适当性检验中的三个争议问题，本研究对国内 9 份管理学、心理学核心期刊 2014 年以来发表的相关文献进行了内容分析和评价，总结了研究中的普遍性问题表现，并提

出初步的解决措施和操作程序。我们不以批评和挑刺为目的，而是希望研究者能够意识到某些习惯性做法的不妥之处并及时补救，力求更可靠、更精确地测量高层次构念。当然，本研究距离彻底解决问题并确立“最佳实践”模式还有很远的距离，很多研究缺口仍有待填补。近期尤其值得关注的是多层次结构方程模型的应用，它将高层次构念按潜变量来建模，对测量误差和抽样误差进行双重校正(毕向阳, 2019)，可实现“潜”聚合，比忽略测量误差而简单取均值的“显”聚合有更高的估计精度，有望改变聚合问题的研究走向。

最后要强调的是，研究者不能仅仅将这一系列检验当做数据驱动下的简单决策过程或论文评审所需的“统计仪式”，而应熟悉背后的原理，增强对理论的关照和审视。多层次研究的一个基本前提是理论、测量和分析必须处于同一层次，否则就会造成研究层次的混淆和谬误(Mathieu & Chen, 2011)，而很多高层次构念的数据只能由团体内个体报告的结果汇总而来，为缓和这种矛盾，必须通过系统的聚合适当性检验和信效度检验来证明低层次数据能够有效代表高层次构念的潜在水平。但统计检验不能代替理论分析，数据聚合在本质上应当是由理论驱动的，逻辑起点是对高层次构念理论合理性的论证。研究者必须对高层次构念为何定位于团体或组织层次、高层次构念的测量方法、高层次构念与实际测量的低层次构念间的关系、推动构念由低层次上升到高层次的团体内互动过程等一系列问题形成周密的思考和清晰的阐释(George & James, 1993; González-Romá, 2019; Morgeson & Hofmann, 1999)，但这是研究者在实践中比较欠缺的。我们在文献梳理过程中发现，对理论问题的轻视已经引发了两个不良后果：一是构念所在的层次混乱，例如有的研究将主管的领导风格聚合到团队层次，有的研究却放在个体层次处理[11]；二是抛开理论设定，单纯根据统计检验结果决定构念的分析层次，例如发现某个理论上应处于团体层次的构念的 ICC 值过低，就将其直接作为个体层次构念纳入后续分析，完全不管这样做是否有道理。为避免层面误设，研究者务必先依据理论确定每一构念（包括控制变量）所在的层次并在文中论述缘由，再采用聚合适当性检验、信效度检验或非独立性检验[12]等统计手段去验证这些设定是否得到数据的支持，而不是任由数据来支配理论的建构。

## 参考文献

毕向阳. (2019). 基于多水平验证性因子分析的城市社区社会资本测量——实例研究及相关方法综述. *社会学研究*, (6), 213–237.

邓今朝, 喻梦琴, 丁栩平. (2018). 员工建言行为对团队创造力的作用机制. *科研管理*, *39*(12), 171–178.

[11] 这一问题暗含了平均领导风格与对偶式（dyadic）/个体差异式领导风格的理论分歧，但多数研究并未言明自己的理论取向。

[12] 对于多层次模型中的个体层次结果变量，应计算 ICC(1)值，较大的 ICC(1)表示个体的评分带有群集性（clustering）或非独立性（non-independence），可能受到团体层次构念的影响(Bliese, 2000)，但无需评估组内一致性（无理论意义）。

方杰，邱皓政，张敏强．(2011)．基于多层结构方程模型的情境效应分析——兼与多层线性模型比较．*心理科学进展*, *19*(2), 284–292.

方杰，张敏强，邱皓政. (2010)．基于阶层线性理论的多层级中介效应．*心理科学进展*, *18*(8), 1329–1338.

韩志伟，刘丽红．(2019)．团队领导组织公民行为的有效性：以双维认同为中介的多层次模型检验．*心理科学*, *42*(1), 137–143.

蒋丽，李永娟，田晓明. (2012)．气氛强度：理论基础及其研究框架．*心理科学*, *35*(6), 1466–1473.

李敏，周恋．(2015)．基于工会直选调节作用的劳动关系氛围、心理契约破裂感知和工会承诺的关系研究．*管理学报*, *12*(3), 364–371.

廖卉，庄瑗嘉．(2012)．多层次理论模型的建立及研究方法．见 陈晓萍，徐淑英，樊景立(编), *组织与管理研究的实证方法(第二版)* (pp. 442–476)．北京：北京大学出版社.

林钲棽，彭台光. (2006)．多层次管理研究：分析层次的概念、理论和方法．*管理学报(台)*, *23*(6), 649–675.

罗胜强，姜嬿. (2014)．*管理学问卷调查研究方法*．重庆：重庆大学出版社.

吕洁，张钢. (2015)．知识异质性对知识型团队创造力的影响机制：基于互动认知的视角．*心理学报*, *47*(4), 533–544.

马君，张昊民，杨涛．(2015)．绩效评价、成就目标导向对团队成员工作创新行为的跨层次影响．*管理工程学报*, *29*(3), 62–71.

田雪垠，郑蝉金，郭少阳，贺冠瑞. (2019)．基于多层验证性因素分析的各种信度系数方法．*心理学探新*, *39*(5), 461–467.

王孟成，毕向阳. (2018)．*潜变量建模与 Mplus 应用 进阶篇*．重庆：重庆大学出版社.

温福星，邱皓政. (2015)．*多层次模式方法论：阶层线性模式的关键问题与试解*．北京：经济管理出版社.

辛自强. (2018)．*心理学研究方法新进展*．北京：北京师范大学出版社.

徐晓锋，刘勇. (2007)．评分者内部一致性的研究和应用．*心理科学*, *30*(5), 1175–1178.

杨建锋，王重鸣. (2008)．类内相关系数的原理及其应用．*心理科学*, *31*(2), 434–437.

于海波，方俐洛，凌文辁. (2004)．组织研究中的多层面问题．*心理科学进展*, *12*(2), 462–471.

张勇，龙立荣，贺伟. (2014)．绩效薪酬对员工突破性创造力和渐进性创造力的影响．*心理学报*, *46*(12), 1880–1896.

张志学. (2010)．组织心理学研究的情境化及多层次理论．*心理学报*, *42*(1), 10–21.

Bartko, J. J. (1976). On various intraclass correlation reliability coefficients. *Psychological Bulletin*, *83*(5), 762–765.

Biemann, T., Cole, M. S., & Voelpel, S. (2012). Within-group agreement: On the use (and misuse) of $r_{WG}$ and $r_{WG(J)}$ in leadership research and some best practice guidelines. *The Leadership Quarterly*, *23*(1), 66–80.

Bliese, P. D. (1998). Group size, ICC values, and group-level correlations: A simulation. *Organizational Research Methods*, *1*(4), 355–373.

Bliese, P. D. (2000). Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis. In K. J. Klein & S. W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations* (pp. 349–381). San Francisco: Jossey-Bass.

Bliese, P. D., Maltarich, M. A., Hendricks, J. L., Hofmann, D. A., & Adler, A. B. (2019). Improving the measurement of group-level constructs by optimizing between-group differentiation. *Journal of Applied Psychology*, *104*(2), 293–302.

Brown, R. D., & Hauenstein, N. M. A. (2005). Interrater agreement reconsidered: An alternative to the $r_{wg}$ indices. *Organizational Research Methods*, *8*(2), 165–184.

Burke, M. J., Cohen, A., Doveh, E., & Smith-Crowe, K. (2018). Central tendency and matched difference approaches for assessing interrater agreement. *Journal of Applied Psychology*, *103*(11), 1198–1229.

Carron, A. V., Brawley, L. R., Eys, M. A., Bray, S., Dorsch, K., Estabrooks, P., et al. (2003). Do individual perceptions of group cohesion reflect shared beliefs? An empirical analysis. *Small Group Research*, *34*(4), 468–496.

Castro, S. L. (2002). Data analytic methods for the analysis of multilevel questions: A comparison of intraclass correlation coefficients, $r_{wg(j)}$ , hierarchical linear modeling, within- and between-analysis, and random group resampling. *The*

*Leadership Quarterly*, *13*(1), 69–93.

Chan, D. (1998). Functional relations among constructs in the same content domain at different levels of analysis: A typology of composition models. *Journal of Applied Psychology*, *83*(2), 234–246.

Cohen, A., Doveh, E., & Eick, U. (2001). Statistical properties of the $r_{WG(J)}$ index of agreement. *Psychological Methods*, *6*(3), 297–310.

Cohen, A., Doveh, E., & Nahum-Shani, I. (2009). Testing agreement for multi-item scales with the indices $r_{WG(J)}$ and $AD_{M(J)}$. *Organizational Research Methods*, *12*(1), 148–164.

Dixon, M. A., & Cunningham, G. B. (2006). Data aggregation in multilevel analysis: A review of conceptual and statistical issues. *Measurement in Physical Education and Exercise Science*, *10*(2), 85–107.

Dunlap, W. P., Burke, M. J., & Smith-Crowe, K. (2003). Accurate tests of statistical significance for $r_{WG}$ and average deviation interrater agreement indexes. *Journal of Applied Psychology*, *88*(2), 356–362.

Dyer, N. G., Hanges, P. J., & Hall, R. J. (2005). Applying multilevel confirmatory factor analysis techniques to the study of leadership. *The Leadership Quarterly*, *16*(1), 149–167.

Farmer, S. M., Van Dyne, L., & Kamdar, D. (2015). The contextualized self: How team-member exchange leads to coworker identification and helping OCB. *Journal of Applied Psychology*, *100*(2), 583–595.

Geldhof, G. J., Preacher, K. J., & Zyphur, M. J. (2014). Reliability estimation in a multilevel confirmatory factor analysis framework. *Psychological Methods*, *19*(1), 72–91.

George, J. M., & James, L. R. (1993). Personality, affect, and behavior in groups revisited: Comment on aggregation, levels of analysis, and a recent application of within and between analysis. *Journal of Applied Psychology*, *78*(5), 798–804.

Glick, W. H. (1985). Conceptualizing and measuring organizational and psychological climate: Pitfalls in multilevel research. *Academy of Management Review*, *10*(3), 601–616.

González-Romá, V. (2019). Three issues in multilevel research. *The Spanish Journal of Psychology*, *22*(e4), 1–7.

James, L. R. (1982). Aggregation bias in estimates of perceptual agreement. *Journal of Applied Psychology*, *67*(2), 219–229.

James, L. R., Demaree, R. G., & Wolf, G. (1984). Estimating within-group interrater reliability with and without response bias. *Journal of Applied Psychology*, *69*(1), 85–98.

James, L. R., Demaree, R. G., & Wolf, G. (1993). $r_{wg}$: An assessment of within-group interrater agreement. *Journal of Applied Psychology*, *78*(2), 306–309.

Jebb, A. T., Tay, L., Ng, V., & Woo, S. (2019). Construct validation in multilevel studies. In S. E. Humphrey & J. M. LeBreton (Eds.), *The handbook of multilevel theory, measurement, and analysis* (pp. 253–278). Washington, DC: American Psychological Association.

Jiang, K., Chuang, C.-H., & Chiao, Y.-C. (2015). Developing collective customer knowledge and service climate: The interaction between service-oriented high-performance work systems and service leadership. *Journal of Applied Psychology*, *100*(4), 1089–1106.

Kirkman, B. L., Tesluk, P. E., & Rosen, B. (2001). Assessing the incremental validity of team consensus ratings over aggregation of individual-level data in predicting team effectiveness. *Personnel Psychology*, *54*(3), 645–667.

Klein, K. J., Conn, A. B., Smith, D. B., & Sorra, J. S. (2001). Is everyone in agreement? An exploration of within-group agreement in employee perceptions of the work environment. *Journal of Applied Psychology*, *86*(1), 3–16.

Klein, K. J., & Kozlowski, S. W. J. (2000). From micro to meso: Critical steps in conceptualizing and conducting multilevel research. *Organizational Research Methods*, *3*(3), 211–236.

Kozlowski, S. W. J., & Hattrup, K. (1992). A disagreement about within-group agreement: Disentangling issues of consistency versus consensus. *Journal of Applied Psychology*, *77*(2), 161–167.

Kozlowski, S. W. J., & Klein, K. J. (2000). A multilevel approach to theory and research in organizations: Contextual,

temporal, and emergent processes. In K. J. Klein & S. W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations* (pp. 3–90). San Francisco: Jossey-Bass.

Krasikova, D. V., & LeBreton, J. M. (2019). Multilevel measurement: Agreement, reliability, and nonindependence. In S. E. Humphrey & J. M. LeBreton (Eds.), *The handbook of multilevel theory, measurement, and analysis* (pp. 279–304). Washington, DC: American Psychological Association.

Lance, C. E., Butts, M. M., & Michels, L. C. (2006). The sources of four commonly reported cutoff criteria: What did they really say? *Organizational Research Methods*, *9*(2), 202–220.

Lang, J. W. B., Bliese, P. D., & de Voogt, A. (2018). Modeling consensus emergence in groups using longitudinal multilevel models. *Personnel Psychology*, *71*(2), 255–281.

Lang, J. W. B., Bliese, P. D., & Runge, J. M. (in press). Detecting consensus emergence in organizational multilevel data: Power simulations. *Organizational Research Methods*. doi: 10.1177/1094428119873950

LeBreton, J. M., James, L. R., & Lindell, M. K. (2005). Recent issues regarding $r_{WG}$, $r^*_{WG}$, $r_{WG(J)}$, and $r^*_{WG(J)}$. *Organizational Research Methods*, *8*(1), 128–138.

LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods*, *11*(4), 815–852.

Lüdtke, O., & Robitzsch, A. (2009). Assessing within-group agreement: A critical examination of a random-group resampling approach. *Organizational Research Methods*, *12*(3), 461–487.

Mathieu, J. E., & Chen, G. (2011). The etiology of the multilevel paradigm in management research. *Journal of Management*, *37*(2), 610–641.

Meyer, R. D., Mumford, T. V., Burrus, C. J., Campion, M. A., & James, L. R. (2014). Selecting null distributions when calculating $r_{wg}$: A tutorial and review. *Organizational Research Methods*, *17*(3), 324–345.

Morgeson, F. P., & Hofmann, D. A. (1999). The structure and function of collective constructs: Implications for multilevel research and theory development. *Academy of Management Review*, *24*(2), 249–265.

Moritz, S. E., & Watson, C. B. (1998). Levels of analysis issues in group psychology: Using efficacy as an example of a multilevel model. *Group Dynamics: Theory, Research, and Practice*, *2*(4), 285–298.

Newman, D. A., & Sin, H.-P. (2020). Within-group agreement ($r_{WG}$): Two theoretical parameters and their estimators. *Organizational Research Methods*, *23*(1), 30–64.

Ng, K.-Y., Koh, C., Ang, S., Kennedy, J. C., & Chan, K.-Y. (2011). Rating leniency and halo in multisource feedback ratings: Testing cultural assumptions of power distance and individualism-collectivism. *Journal of Applied Psychology*, *96*(5), 1033–1044.

O'Neill, T. A. (2017). An overview of interrater agreement on Likert scales for researchers and practitioners. *Frontiers in Psychology*, *8*. doi: 10.3389/fpsyg.2017.00777

Quigley, N. R., Tekleab, A. G., & Tesluk, P. E. (2007). Comparing consensus- and aggregation-based methods of measuring team-level variables: The role of relationship conflict and conflict management processes. *Organizational Research Methods*, *10*(4), 589–608.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Newbury Park, USA: Sage.

Rego, A., Cunha, M. P., & Simpson, A. V. (2018). The perceived impact of leaders' humility on team effectiveness: An empirical study. *Journal of Business Ethics*, *148*(1), 205–218.

Schaubroeck, J. M., Shen, Y., & Chong, S. (2017). A dual-stage moderated mediation model linking authoritarian leadership to follower outcomes. *Journal of Applied Psychology*, *102*(2), 203–214.

Schneider, B., White, S. S., & Paul, M. C. (1998). Linking service climate and customer perceptions of service quality: Tests

of a causal model. *Journal of Applied Psychology*, *83*(2), 150–163.

Shen, J. (2016). Principles and applications of multilevel modeling in human resource management research. *Human Resource Management*, *55*(6), 951–965.

Smith-Crowe, K., Burke, M. J., Cohen, A., & Doveh, E. (2014). Statistical significance criteria for the $r_{WG}$ and average deviation interrater agreement indices. *Journal of Applied Psychology*, *99*(2), 239–261.

Smith-Crowe, K., Burke, M. J., Kouchaki, M., & Signal, S. M. (2013). Assessing interrater agreement via the average deviation index given a variety of theoretical and methodological problems. *Organizational Research Methods*, *16*(1), 127–151.

Van Mierlo, H., Vermunt, J. K., & Rutte, C. G. (2009). Composing group-level constructs from individual-level survey data. *Organizational Research Methods*, *12*(2), 368–392.

Woehr, D. J., Loignon, A. C., Schmidt, P. B., Loughry, M. L., & Ohland, M. W. (2015). Justifying aggregation with consensus-based constructs: A review and examination of cutoff values for common aggregation indices. *Organizational Research Methods*, *18*(4), 704–737.



**附录：部分计算工具**

（1）LeBreton 和 Senter (2008)给出了在多种虚无分布形态和不同计分点数下的期望变异值（pp.832-833），在计算 $r_{WG}$ 时可参考。

（2）Biemann 等(2012)开发了基于 Excel 软件的免费小工具（访问 www.sbuweb.tcu.edu/mcole 下载），可以容易地计算 $r_{WG}$、ICC(1)、ICC(2)、*F* 值等，并给出了展示聚合适当性检验结果的表格模板（p.78）。

（3）Krasikova 和 LeBreton (2019)编写了计算 $r_{WG}$、ICC(1)、ICC(2)的 R 软件代码（pp.300-302）。

（4）温福星和邱皓政(2015)给出了用 SPSS 计算 $r_{WG}$ 的语法示例（pp.55-57）。

（5）罗胜强和姜嬿(2014)给出了用 SPSS 计算 ICC(1)和 ICC(2)的示例（pp.280-283）。

# Data aggregation adequacy testing in multilevel research: A critical literature review and preliminary solutions to key issues

ZHU Haiteng

(*Department of Military and Ideological Basic Education, PLA Army Academy of Artillery and Air Defense, Hefei* 230031, *China*)

**Abstract:** The measurement of shared unit property constructs is ubiquitous in multilevel organizational research, of which the most frequently used approach is to aggregate the ratings of several unit members to the unit level. The data aggregation adequacy testing (DAAT) is a statistical hurdle to ensure the validity and representativeness of aggregated scores. Well-established indicators of DAAT include within-group agreement index, $r_{WG}$, and within-group reliability indices, ICC(1) and ICC(2); nonetheless, some key issues are still open to debate, for instance, the superiority of the two families of indicators, the null distribution and data screening decision of $r_{WG}$, and appropriate cut-off values. To address the above questions, the current

research firstly conducted a content analysis of 166 studies adopting DAAT procedure published on 9 Chinese journals in the field of management and psychology since 2014, coupled with 85 studies from *Journal of Applied Psychology* as a comparison. Common problems in routine practice of DAAT were identified and related suggestions were proposed as follows: (1) Disentangling and differentiating the role of DAAT indicators; specifically, $r_{WG}$ should be used as the exclusive indicator of aggregation adequacy, whereas ICC(1) and ICC(2) should be deemed as indices of validity and reliability, respectively. (2) Making prudent and justifiable decisions in choosing null distributions when calculating $r_{WG}$ index, and excluding groups with low within-group agreement. (3) Applying more reasonable and moderately flexible cut-off values instead of arbitrary and rough practical standards. Last but not the least, researchers should always prioritize theoretical considerations in the process of framework building and DAAT, and unload disproportionate dependence on statistical results.
**Key words:** multilevel research; shared unit property; aggregation; within-group agreement; within-group reliability



**作者贡献申明：**本文作者独立完成了提出研究问题、设计研究框架、分析研究数据、写作和修改论文等工作；文献检索和编码工作得到姚小雪博士的支持，在此谨致谢意。